Abstract

For the task of near-duplicate document detection, comparison approaches based on bag-of-words used in information retrieval community are not sufficiently accurate. This work presents novel approach when instance-level constraints are given for documents and it is needed to retrieve them, given new query document for near-duplicate detection. The framework incorporates instance-level constraints and clusters documents into groups using novel clustering approach Grouped Latent Dirichlet Allocation (gLDA). Then distance metric is learned for each cluster using large margin nearest neighbor algorithm and finally ranked documents for given new unknown document using learnt distance metrics. The variety of experimental results on various datasets demonstrate that our clustering method (gLDA with side constraints) performs better than other clustering methods and the overall approach outperforms other near-duplicate detection algorithms.